Electoral Campaigns and Relation Mining: Extracting Semantic Network Data from Newspaper Articles

Wüest, Bruno ; Clematide, S ; Bünzli, A ; Laupper, D ; Frey, T

Abstract: Among the many applications in social science for the entry and management of data, there are only a few software packages that apply natural language processing to identify semantic concepts such as issue categories or political statements by actors. Although these procedures usually allow efficient data collection, most have difficulty in achieving sufficient accuracy because of the high complexity and mutual relationships of the variables used in the social sciences. To address these flaws, we suggest a (semi-) automatic annotation approach that implements an innovative coding method (Core Sentence Analysis) by computational linguistic techniques (mainly entity recognition, concept identification, and dependency parsing). Although such computational linguistic tools have been readily available for quite a long time, social scientists have made astonishingly little use of them. The principal aim of this article is to gather data on party-issue relationships from newspaper articles. In the first stage, we try to recognize relations between parties and issues with a fully automated system. This recognition is extensively tested against manually annotated data of the coverage in the boulevard newspaper Blick of the Swiss national parliamentary elections of 2003 and 2007. In the second stage, we discuss possibilities for extending our approach, such as by enriching these relations with directional measures indicating their polarity.

DOI: https://doi.org/10.1080/19331681.2011.567387

First Submission: 06/07/2010
Revised Submission: 02/04/2011
Accepted: 02/04/2011

Electoral Campaigns and Relation Mining: Extracting Semantic Network Data from Newspaper Articles

Bruno Wueest, Simon Clematide, Alexandra Bünzli, Daniel Laupper and Timotheos Frey

University of Zurich

Abstract

Among the many applications in social science for the entry and management of data, there are only a few software packages that apply natural language processing to identify semantic concepts such as issue categories or political statements by actors. Although these procedures usually allow efficient data collection, most have difficulty in achieving sufficient accuracy because of the high complexity and mutual relationships of the variables used in the social sciences. To address these flaws, we suggest a (semi-)automatic annotation approach that implements an innovative coding method (Core Sentence Analysis) by computational linguistic techniques (mainly entity recognition, concept identification, and dependency parsing). Although such computational linguistic tools have been readily available for quite a long time, social scientists have made astonishingly little use of them. The principal aim of this paper is to gather data on party-issue relationships from newspaper articles. In the first stage, we try to recognize relations between parties and issues with a fully automated system. This recognition is extensively tested against manually annotated data of the coverage in the boulevard newspaper *Blick* of the Swiss national parliamentary elections of 2003 and 2007. In the second stage, we discuss possibilities for extending our approach, such as by enriching these relations with directional measures indicating their polarity.

For decades, scholars in the social sciences have been coding written documents manually to gain empirical insights. However, such procedures entail various problems. The most severe drawback of manual coding – especially of large-scale data collections – is the enormous time and expense required (Schrodt, 2009; Hillard, Purpura, & Wilkerson, 2007). At the same time, the quality of the data gathered often is difficult to assess, mainly as a result of rather low inter- and intra-coder reliabilities[1] or problematic construct validities of theoretically complex variables. The flaws of traditional coding methods are becoming ever more severe in the context of "big data," i.e., the increasing availability of documents that are interesting for scientific research. The pressure on political institutions to become more transparent and the growth of the Internet have led to a swelling proliferation of digitally available documents from various sources (Cortada, 2009; Fung, Graham, & Weil, 2007). At the same time, the sheer amount of data makes those documents increasingly inaccessible.

The dominant paradigm in political science to handle "big data" is the estimation of semantic information in documents by means of statistical procedures (Hopkins & King, 2010; Laver, Benoit, & Garry, 2003; Zuell & Landmann, 2005; Hillard et al., 2007).[2] A

[1] *Reliability* is concerned with the question of whether the data collection is more stable over time, better reproducible by different coders, and more accurate compared to some canonical standard (see Krippendorff, 2004). Inter- and intra-coder agreement are common terms in social science content analysis. In linguistics, in contrast, the terms inter-annotator agreement and consistency are used.

[2] There is another interesting line of applied computational linguistics in political science. The Kansas Event Data System (KEDS) (Schrodt, Davis, & Weddle, 1994), its successor, TABARI, and the further developed version, called VRA-Reader (King & Lowe, 2003), all apply relation-oriented procedures to identify event data. These programs process newswire leads and search for sources, actions, and targets, which are stored in large dictionaries. The most serious drawback of such software is that it depends on the highly standardized language of press reports because it is not able to parse natural language. For continuously updated information, see Neuendorf & Skalski (2010).

common feature of these approaches is that they code articles using one specific variable, be it the left-right scale, issue categories, or ordinal variables. Such procedures rely on the comparison of relative word frequencies (Laver et al., 2003; Zuell & Landmann, 2005; Hillard et al., 2007), the co-occurrence of a few keywords (Ruigrok & van Atteveldt, 2007), or dichotomous variables assessing the presence of word stems (Hopkins & King, 2010).

One impressive advantage of these approaches is their independence from costly resources such as large keyword dictionaries. All these procedures need is a sufficient amount of manually coded reference texts, usually not more than 100 documents, and they basically handle all kinds of unstructured texts (Hopkins & King, 2010). However, this doesn't hold without restrictions. Changes of vocabulary over time and between different authors as well as large differences in the length of the documents can lead to imprecise coding assessments (Hug & Schulz, 2007).

The decisive disadvantage for us, however, is that these methods do not meet our primary research goal: the recognition of relational data.[3] First, some approaches simply do not aim at linking issues to specific actors but try to classify texts (e.g. Hillard et al., 2007). Second, approaches aiming at generating relational data have mainly been used to code party manifestos, parliamentary speeches, or weblogs for which the actors are already pre-defined. However, the simultaneous occurrence of actors and issues, as well as contradicting political positions of the same actor, is not rare in newspapers. For example, an article may report the electoral campaign efforts of several parties. Another article may, instead, focus on one party but discuss deviating statements of its exponents regarding a single policy.

[3] By relational data, we mean a tuple consisting of a political actor (e.g., a party), an issue, and the polarity of the relationship between actor and issue. For the implications of this definition for the manual content analysis, see the next section on our measure of party competition. For the presentation of the actual relational data, see Figure 3 on page 13.

To gather relational data from newspaper articles, we make use of advancements in the field of computational linguistics, where the interest in information extraction and relation detection has considerably grown in recent years (Jurafsky & Martin, 2008; Scharl & Weichselbraun, 2008; Rinaldi et al., 2007; Porter, Barker, & Hovy, 2007; West, 2001; Evans, 2001). While social scientists, so far, have made surprisingly little effort to integrate approaches such as named entity recognition,[4] concept identification, and syntactic analysis into their content analyses, abundant computational linguistic work has shown that these techniques can help to find relations between specific concepts (see van Atteveldt, Kleinnijenhuis, & Ruigrok, 2008). In the following, we will outline our conceptual framework, discuss the technical and linguistic implementation, and present an evaluation of our approach.

An Integral Measure of Party Competition

Our starting point is the improvement of an innovative data collection method for measuring party positions.[5] This approach, Core Sentence Analysis (CSA), has its origins in early theoretical elaborations by Wittgenstein (1984 [1921]) and was first implemented into concrete coding instructions by Osgood (1956) and Axelrod (1976). Recently, it has been renewed for the analysis of electoral campaigns and political conflict in general (Kriesi et al., 2008; Kleinnijenhuis, de Ridder, & Rietberg, 1997). Additionally, Franzosi (2004) provided theoretical and empirical evidence that the method – he called it "story grammars" – is a useful device for the social sciences in general. It is an inductive approach to capture the full

[4] *Named entity recognition* means the detection of proper names, e.g. names of parties or politicians, in text documents. Assigning different mentions of proper names in a text to the same reference entity is called alias resolution.

[5] For a comprehensive description of the manual data collection and analysis, see Kriesi et al. (2008).

complexity of the political debate without imposing theoretical expectations on the data, which constitutes a common problem for content analysis.

The basic idea of this method is that the content of every written document can be described as a network of objects. In our case, we identify relationships between "political objects," i.e. between a political actor and a political issue (see Table 1). Each sentence of a document is reduced to its most basic semantic structure (the so-called core sentence), consisting of a subject (actor which is either a party or a politician), its object (issue), and the direction of the relationship (polarity) between the two (using a scale ranging from -1 to +1 with three intermediary positions).

(Insert Table 1 about here)

If the actors and issues are aggregated into meaningful categories, an election campaign can be mapped and evaluated by constructing a network of positions and saliencies.[6] The position is calculated by taking the average of all statements of one party regarding a specific issue, while the saliencies indicate the relative frequency with which a party positions itself on this specific issue.

If the data are collected from several hundred newspaper articles of different newspapers, they allow us to investigate the party programs more precisely than is possible in more common data sets, such as party manifesto data, expert surveys, or roll call data (Kriesi et al., 2008). Moreover, such data can be used in state-of-the-art research on party competition (e.g. Adams, Merrill, & Grofman, 2005; Meguid, 2005) that combines positions as used in strategic models (e.g. Downs, 1957 or Rabinowitz & MacDonald, 1989)[7] with the

[6] Our actor aggregations are simply the parties competing for the national parliament; our issue categorization can be found in Table A.1 in the appendix.

[7] Directional or proximity spatial models.

concept of relative frequencies[8] employed by theories of selective emphasis and issue ownership (Budge & Farlie, 1983; Petrocik, 2004).

In the following, we will present the technical implementation of our method to automate the coding task as much as possible. At least for the moment, however, the automatic recognition of the polarity of relationships is too difficult and, therefore, out of reach for the present paper. Instead, we will first focus on the computer-assisted detection of relationships and try to gather data on relative frequencies. Thereafter, we will discuss future prospects on how polarity can be assessed automatically.

Implementing the Core Sentence Analysis Approach

For the processing and coding of the articles, we used and further developed various linguistic tools. Figure 1 presents an overview of our text analysis pipeline that integrates the different processing steps by means of a standardized XML format; i.e. throughout the processing, an XML file containing the intermediate results of each step can be produced for controls or further analysis.[9]

(Insert Figure 1 about here)

Newspaper articles serve as input to the pipeline, while the output is the processed article with its metadata, the annotated concepts, and the recognized relations. The pipeline is designed in a modular way in order to process different languages and use different tools. Therefore, it is possible to adapt to other languages without too much effort. For this paper, however, only a German version was applied.

[8] In political science, relative frequencies, in this context, are usually denoted as salience. Since we do not want to confound this meaning with other and, especially, linguistic denotations, we will stick to the more specific term "relative frequency."

[9] *XML* (Extensible Markup Language) is a standard format for encoding documents as structured textual data in machine-readable form.

Standardization

First, the articles are converted and normalized into our pre-defined XML representation since they originate from different sources such as CD-ROMs or digital archives of publishing houses. Important metadata such as the newspaper title, publishing date, length, rubric, and author is encoded in a uniform manner.

Text Segmentation and Tokenizing[10]

We adapted the tokenizer developed at the University of Stuttgart by Schmid (1994) to split a running text into tokens and sentences. This step generates a unique ID for every token and sentence. For example, the sentence "*Die FDP ist ohne Wenn und Aber für Steuersenkungen*" results in the following structure:

<sentence id="bli_2003_10_04_3965-61">

*<token id="bli_2003_10_04_3965-61-1">**Die**</token>*

*<token id="bli_2003_10_04_3965-61-2">**FDP**</token>*

*<token id="bli_2003_10_04_3965-61-3">**ist**</token>*

[...]

</sentence>,

where "*bli_2003_10_04_3965*" is the article identifier.

Part-of-Speech[11] *Tagging*[12] *and Morphological Analysis*

To perform concept identification in German, the base forms of the inflected words (the *lemma*) are needed.[13] For this lemmatization step, we apply the morphological analysis tool

[10] Tokens are a sequence of characters that serve as basic elements for further linguistic processing. Typical tokens are word forms and punctuation marks.

[11] *Part-of-speech* is the lexical category of a word. The most important lexical categories are nouns, verbs, adjectives, and prepositions.

[12] A tagger marks every token (words and punctuation marks) with a part-of-speech label, a so-called *tag*.

GERTWOL (Koskeniemmi & Haapalainen, 1996). To increase precision, the words are first tagged with their part-of-speech labels by the TnT-Tagger (Brants, 2000). For short text segments consisting of 4 tokens or fewer, this is the only syntactical analysis step we perform.

Named Entity Recognition (NER) and Concept Identification

The next step is the recognition of the politicians and issues of interest. We chose an approach with carefully handcrafted lists because the task specification demands high accuracy. The gazetteer[14] contains 2710 persons with information about their party affiliation, gender, and VIP status, i.e. whether they are famous politicians. In contrast to less-prominent persons, VIPs are often referred to only by their last names. There are about 1990 different last names on our list, and frequent ones, such as Müller, refer to as many as 30 different politicians. The alias resolution, i.e. the recognition of actors, is done at the level of the whole article: first, every occurrence of a single last name (in the case of VIPs) or a combination of a first and last name (in all other cases) is matched to the entries in our list. Second, journalists often mention the full name of politicians once at the beginning and then refer to this politician only by her or his last name.[15] Therefore, all single last name mentions without VIP status for which we found a combination of a first and a last name in other sections of

---

[13] For example, "go" is the lemma of "went," "goes," or "gone."

[14] *Gazetteer* is a list of names for specific concepts held in computer form, which allows for rapid search and query.

[15] The following example may be exemplary for this case *(Blick*, August 22, 2003): "'Es ist noch nicht zu spät', ist SP-Nationalrat Rudolf Strahm (60) überzeugt. [...] 5 Millionen Franken müssten laut Strahm reichen, um die heute noch stellenlosen Jugendlichen unterzubringen [...]"; [*"SP-National Councillor Rudolf Strahm (60) is convinced that ‚it is not too late'. [...] According to Strahm, 5 billion Swiss Francs should be enough to get the currently unemployed young a job."*]

the article are also attributed to this politician. Other detected mentions of single last names are discarded.

Another task in NER is anaphora resolution.[16] After a first mention by his or her name, a politician is often referred to using pronouns or definite noun phrases. State-of-the-art methods for anaphora resolution in German were established by Wunsch (2010) and Klenner and Ailloud (2009). Since the accuracy of the resolution of personal pronouns is still limited (e.g., the hybrid system developed by Wunsch (2010) performs at an F-measure of 55%), we restricted the coreference resolution to a very straightforward procedure: If we encounter the personal pronouns "sie" (she) or "er" (he) in any grammatical case, and, at the same time, a politician of the corresponding gender is found in the previous sentence, the pronoun is resolved to this politician. Additionally, we treat cases where "sie" refers to parties (in the meaning of "it"), as well as uses of "wir" (we), as references to the collective actor (e.g. "we, the party...") that occur quite often in interviews.

The identification of political issues requires slightly different methods than the recognition of politicians and parties. The issue gazetteer contains a list of manually built trigger patterns for each issue concept. In the simplest case, this is just the base form of a single word, e.g. the compound "Steuersenkung" (tax reduction). However, more often, one word in isolation is usually too general or too ambiguous for reliable concept identification. Hence, our trigger patterns allow for Boolean combinations of further keywords that must appear in the same sentence. Including automatically generated orthographic variants, more than 2100 trigger patterns have been defined. 1288 of them consist of a single keyword, 791 include two keywords, and 74 are made of three or more keywords.

[16] *Anaphora* are linguistic elements, mostly pronouns or definite nominal phrases, that refer to other elements in a text.

The concept identification algorithm works with a longest-pattern-first strategy. Further, some trigger words[17] are ambiguous with regard to our issue categories (see Table A.1 in the appendix). For instance, "Drogenpolitik" (drug policy) may trigger for the category of either cultural liberalism or security. Such ambiguities are resolved on the document level by selecting the issue category with the maximum number of unambiguous hits within the same document. There were three ambiguous keywords triggering an ambiguity resolution in only 29 cases. Of these 29, 21 (72%) were coded correctly, which is a good achievement. However, considering the fact that 29 cases is a small number, this specific rule did not have a big impact on the results.

Syntactic Analysis[18]

The weighted constraint dependency grammar (CDG) parser (Foth, Menzel, & Schröder, 2005) is used for a robust syntactic analysis.[19] Its broad coverage dependency grammar for German has been evaluated on newspaper texts and 87% of all labelled dependencies can be expected to be correct (Foth, Daum, & Menzel, 2004). At the time of our project start, to our knowledge, no other publicly available German parser could compete with this tool. Although the parser already has a large lexicon, it had to be further adapted to the specific vocabulary (especially nouns and verbs) of Swiss politics and Swiss orthography. Unlike other statistical parsers, even grammaticality conditions can be adapted to the domain. One beneficial feature of the parser is that it produces dependency trees, which display the

[17] *Trigger patterns* are understood as words in the documents that initiate a computational linguistic procedure, e.g. the recognition of concepts.

[18] *Syntactic dependency* indicates the syntactic relations between phrases (subject, object, predicate, etc.) in a sentence.

[19] *Parsers* compute the syntactical structure of sentences, i.e. which tokens form the subject, verb, objects, and so forth, of the sentence.

argument structures more directly than phrase structure trees. Another useful property of the parser is that it tries hard to deliver the most probable syntactic structure of each sentence. Computing the optimal structure for a sentence can take time exponential to the sentence length in general, and, therefore, using heuristic search strategies is inevitable. Nevertheless, parsing long sentences may take several minutes. For this reason, we set a time limit of 90 seconds for each sentence to parse our articles in an acceptable period. This means that, after the expiration of the time limit, the best analysis found so far is returned. The flipside of this setup is that the parsing of very long sentences delivers suboptimal results, so we have decided not to analyze sentences with more than 35 words.[20] The number of long sentences varies considerably between different newspapers; in the case of the boulevard newspaper *Blick* used for our evaluation, roughly every 150th sentence was affected. To easily examine the quality of the computed syntactic structures, we render the parse trees as scalable vector graphics[21] (see Figure 2a for our example sentence).

(Insert Figure 2.A about here)

(Insert Figure 2.B about here)

Deepening the Syntactic Analysis

After computing the primary syntactic dependencies, we insert secondary relations into the parse tree to shorten the paths between the content-bearing elements. The resulting structures, which resemble predicate-argument-structures, thereby facilitate the task of extracting core sentences. Figure 2b shows a typical dependency tree with such secondary relations: one secondary relation between the "SVP" and the main verb "ablehnen" (reject)

[20] Various solutions are available to partially remedy this problem, e.g. splitting long sentences into subclauses before parsing, or integrating the results of a fast statistical parser into the CDG system (Foth, 2007).

[21] The Perl package DepSVG was developed by K. Kaljurand: information and download at http://files.ifi.uzh.ch/cl/kalju/download/depsvg (April 30, 2010).

and another between the coordinated "FDP" and the main verb. We also deal with other syntactical phenomena, e.g. passive voice or modal verbs. This way, we have the necessary syntactical information needed for the extraction of relations between political actors and issues. Ideally, the desired core sentences would have the form *actor/predicate/issue* and would correspond, more or less, to the syntactic dependencies (see Table 1). In the next section, we present the results of several extraction methods that make use of this information.

Determining the Validity of Our Approach

We chose the two most recent Swiss national parliamentary election campaigns, in 2003 and 2007, to evaluate the validity of our approach. Validity, thus, refers to the question of whether the collected data actually measures the theoretically derived concepts. More specifically, we consider the election coverage in the boulevard newspaper *Blick*, the largest non-free daily newspaper in Switzerland.[22] This decision was, one the one hand, motivated by the consideration that we needed a German-speaking country to evaluate our language-dependent software. Switzerland was then selected because we are most familiar with this country. This facilitates both the development of gazetteers and linguistic rules as well as the interpretation of the results. One the other hand, we have manually annotated data that has been used for actual research at our disposal (e.g. Kriesi et al., 2008; Helbling et al., 2010). This data serves as the gold standard against which the automatically coded data can be compared.[23]

[22] The time period is ten weeks before the polling day for both elections, which we consider the peak period of electoral campaigns.

[23] *Gold standard* (in content analysis evaluations) denotes data generated by methods other than the evaluated procedure. This data is treated as paragon of excellence against which the new data is compared.

To enable the improvement of our method during the evaluation process as well as to ensure precise error analysis, we split all articles into a development set (187 articles) and a test set (90 articles).[24] The development set was used to train our computational tools and linguistic rules as well as to refine our lists of politicians, parties, and issue trigger patterns. The test set was evaluated only once, at the end of the development phase. The quality of the test set codings, thus, serves as an unbiased benchmark of our method since it is applied to previously untreated data. In addition to this general evaluation setting, two coders annotated the test data set for the 2003 election campaign in parallel. With this data at hand, we are in a position to directly assess the agreement among human coders and between the human coders and the machine coding.

We determine annotation validity on the article level since it is often unclear to which sentence a political statement belongs. Especially in the context of anaphora and long quotations of political actors, it is difficult to pick one sentence as a relation's source (see van Atteveldt et al., 2008, p. 436). Accordingly, the manual data in the gold standard often are very imprecise regarding the exact source of a relationship.[25]

In the first step of the evaluation, we will assess the reliability of the automated methods for the actor, issue, and relation recognition separately. Every concept found is

[24] The number of articles is relatively small compared to similar experimental designs. However, as the evaluations will show, the sample is large enough to provide a valid test of our approach.

[25] The following example nicely illustrates this (*Blick*, August 27, 2003): "Anders sein absehbarer Nachfolger, der Thurgauer Bahnunternehmer Peter Spuhler. Er springt in die Lücke und will für alle Betriebe, die mehr als 2 Prozent Lehrlinge beschäftigen, einen Steuerabzug gewähren." [*"Otherwise* [thinks] *his conceivable successor, the rolling stock entrepreneur Peter Spuhler from the canton of Thurgau. He steps into the breach, since wants to allow a tax reduction for all enterprises that employ more than 2 percentages of trainees."*] Here, the recognizable actor of the relation (Peter Spuhler) is introduced one sentence before the issue (tax reduction) is named. Such sentences are rather common in the newspaper articles under study.

assigned to one of three categories: true positives (tp) are cases that are recognized in both the gold standard and the machine-coded data set; false negatives (fn) are cases that are a part of the gold standard but not recognized by the machine coding; and false positives (fp), in contrast, are assigned by the machine coding but not identified in the gold standard. From the frequencies of these categories, we can compute the recall (r), precision (p), and F-score (F1) (see Manning & Schütze, 2002):

$$r = \frac{tp}{tp + fn} \qquad p = \frac{tp}{tp + fp} \qquad F_1 = \frac{2 \times p \times r}{p + r}$$

The recall indicates how often a concept found by the manual annotation was also found by the automated method. In contrast, the precision indicates how often the automated method is right when it recognizes an issue, actor, or relation. The F-score is the harmonic mean of recall and precision; i.e., it collapses the two indicators to a general measure of fit by giving both indicators the same weight.

After a presentation of these measures for different machine-coding methods, we will qualify the accuracy of the automated annotation by comparing it with simple correlation measures to different manually collected data sets. This means that we cease treating the manual annotated data as an unquestionable standard of data collection and try to assess how well the automated annotation works when we control for the inconsistencies of the manual method. At the same time, the external validity of the Core Sentence Analysis (CSA) approach can be established, i.e. how well the CSA annotation can be transferred between different implementations. Finally, the evaluation is deepened by an error components analysis, which estimates the usefulness of the different steps in our coding method.

Recognition Methods

Since the main interest of this paper is how we should best apply our implementation of the CSA approach, we will pit different recognition methods against each other to discuss the

advantages and drawbacks of each method. In the first step, we evaluate how the methods for the recognition of actors and issues perform. Then we will proceed to the evaluation of the relation recognition. Additionally, we will present the evaluation separately for the frequencies and single occurrences of the different concepts per article. The approach of evaluating only single occurrences will increase the precision of our recognition methods since it is easier to establish, if a concept occurs in an article at all, than to exactly identify how often this concept occurs. Since we are, however, interested in data that is as fine-grained as possible, we will also assess the quality of our methods with respect to the frequency per article. Furthermore, for both the actor/issue and relation recognition, we define a baseline system that provides us with a benchmark for the performance of the other methods.

For actor and issue recognition, the baseline system is, simply, all actors and issues that are recognized in an article. Against this most basic method, the following two approaches are pitted. The first is a method, denoted in the following as (*±1s*), that considers only actors and issues that are recognized within a window of 3 sentences of each other. If actors and issues are quite close to each other, they probably are also relevant for the establishment of relations. The second approach (referred to as *sentence* in the following) is similar but uses the sentence boundaries as the window in which actors and issues must jointly occur. The baseline system of the actor-issue relation recognition again is established in a very simple way: all possible pairs of parties and issues that occur in the same sentence are taken. Such an approach is straightforward but neglects various aspects of the relation between parties and issues, which will be tested sequentially by more sophisticated recognition approaches.

On the one hand, the nature of political statements, as relevant for the CSA annotation, is that one actor mostly relates to only one issue and that relations sometimes span more

than one sentence. Therefore, the first improvement will be to apply a method that attaches only the nearest issue to each recognized party. The nearness is measured in terms of token distance and limited to the preceding and following sentences. This method is denoted as (*simple distance*). On the other hand, the baseline system does not pay attention to the syntactical structure of the sentences. An intuitive way to improve the relation detection by considering the syntactical structure is to choose only concepts that are important in terms of their part-of-speech. The third method, denoted as *subject filter,* therefore applies the following rule: either the party or the issue must occur as part of a subject dependency. The last approach, called *parse tree distance*, is our most sophisticated attempt by using the following measure: for each occurrence of parties and issues in a sentence, the pair with the minimal distance in terms of dependency links (minimal dependency path) was selected. In Figure 2, for example, the actor "FDP" has a minimal distance of two dependency links to the issue "Steuersenkungen" if we use the secondary relation. Additionally, a threshold of a maximum of 4 dependency links between the entities was applied with the aim of improving precision.

Performance of the Actor and Issue Recognition

We start with an evaluation of party and issue recognition, since the relation detection essentially relies the a proper identification of these basic units. Table 2 shows the recall, precision, and F-score for the frequencies of these concepts per article. The first column (*article*) indicates all concepts recognized in the same article; for the calculation shown in the second column (*±1*), we included only issues and actors located within 3 sentences from each other; and the third column (*sentence*) presents the strictest criterion: only actors that occur with an issue in the same sentence (and vice versa) are considered here. Additionally, the number of observations (N) shows the sum of entities recognized by these three methods, and the number of observations in the gold standards is indicated by N_G.

(Insert Table 2 about here)

The recall for all approaches obviously is lower in the test set than in the development set, where we had the chance to improve the recognition previous to the evaluation. The precision, however, is surprisingly higher for the test set data. The underlying cause for this is that there are articles that contain especially difficult passages such as enumerations and tables, for which our system performs quite badly. Unfortunately, articles with such passages are more frequent in the development set than in the test set. However, this is not a valid argument that our experiment is not representative since these passages occur regularly in newspaper articles.

Further, the precision mainly goes up as the unit of measurement is narrowed to the sentence level, while the recall decreases. Of course, this is a consequence of the fact that the ±1 sentences and sentence methods exclude a lot of entities. In the development set, the numbers fall from 2346 to 824 and 359 for the actors and from 1458 to 699 and 350 for the issues. The best F-scores for both the parties and the issues are reached by the ±1 sentences method: 0.55 for the parties and 0.58 issues in the development set and 0.56 (parties) and 0.53 (issues) in the test set. The definition of a text passage that spans a few sentences as the unit of measurement, thus, seems to offer the best balance between recall and precision.

One way to improve the quality of the automated actor and issue recognition is to constrain the analysis to the single occurrence per article instead of measuring frequencies. It is easier for our system (and probably most other systems, too) to recognize that a category of a concept occurs than to assess exactly how often this category is present in an article. Therefore, the evaluation measures most likely can improve. Table 3, accordingly, indicates the same measures for the issue and party occurrence. As the figures clearly indicate, this approach is promising since all measures have increased. This finding is intuitively understandable: it is easier to recognize that a concept somehow appears in a document than to identify the exact the number of appearances of this concept. The F-scores for our

preferred method (±1 sentences) increase to 0.74 for the parties and 0.69 for the issues in the development set and to 0.73 (parties) and 0.66 (issues) in the test set. With respect to the development set, this is an improvement of 19 percent for the actors and 11 percent for the issues. Further, these results are far better than those that similar studies involving other languages have achieved (e.g. van Atteveldt et al., 2008).[26]

(Insert Table 3 about here)

Performance of the Relation Recognition

Let us now turn to the relation detection. Again, we present our measures separately for the frequency of relations (Table 4) and their simple occurrence (Table 5) in the documents. The first column in the upper half of the tables (*all combinations*) indicates the most intuitive approach to establish relations by simply combining each party with every issue category in the same sentence. The second column shows the results for the method with the widened scope to detect relations (*simple distance*): every identified issue is combined with its nearest party category. The lower half of the tables further indicates how the two relation detection approaches perform taking syntactical information into account (*subject filter and parse tree distance*). Again, N_G indicates the number of observations in the gold standard.

(Insert Table 4 about here)

The performance of our first method, the detection of all possible party-issue combinations within the same sentence, is hampered by a low recall (e.g. only 0.23 in the test set). Since the two parser-driven methods are equally dependent on the entities found in the

[26] Although comparisons between applications are difficult, since the aim of the analyses, the original data sources and the applied software mostly vary, the study of van Atteveldt et al. (2008) is somewhat comparable since they use a similar software pipeline, process newspaper articles (although in Dutch), and aim to collect CSA data.

same sentence, these problems are also acute here. Furthermore, the longest sentences could not be parsed.[27] Even though taking syntactical information into account helps to improve the precision from 0.59 to 0.63 in the development set and from 0.68 to 0.69 in the test set, respectively, recall decreases to very low levels. Therefore, the *simple distance* method yields the best overall performance with F-scores of 0.50 for the frequencies of relations per article in the development set and 0.59 for occurrences per article (see Table 5). This method widens the scope of the relation detection to the nearest neighbors while ignoring syntactical information. In sum, these superior results are due to the marked increase in recall at a comparatively lower loss of precision. However, if we consider precision to be most important quality, the *simple distance* method fares worst. Accounting for syntactical information in the subject filter method gives consistently the best results with respect to precision.

(Insert Table 5 about here)

Deepening the Evaluation

Table 6 shows the direct comparison of the automated annotation with two manually annotated data sets stemming from the same articles, i.e. the 2003 test set.[28] These two manually annotated data sets are denoted as coder A and coder B, respectively, in Table 6. This allows for the assessment of how well the automated approach works in comparison with the validity of the manual annotation. We took the best-performing methods as described above, which are the *±1 sentence* approach for the parties and issues and the *simple distance* approach for the relations. The columns show the F-scores and Pearson's R.[29]

[27] Here, the application of a more efficient parser (e.g. Sennrich et al., 2009) may alleviate this problem.

[28] These two data sets were also annotated using the CSA approach as described in the section titled *An integral measure of party competition.*

[29] Pearson's R additionally considers true negatives: that is entities that are neither recognized by the gold

(Insert Table 6 about here)

In general, the manual annotation is more reliable than the automated method. Yet again, the F-scores for the comparisons of the automated method with human coders are acceptable. The R scores on the article level further show that the two human coders have very good agreement. The CSA approach is, thus, well-transferable between different human coders. The correlations between the automated data and the human coders are rather high, except, again, for the relation detection. However, this changes when we move to the level of analysis, which usually is a whole election campaign. The last column shows the correlations for the whole test data set of the 2003 electoral campaign. Here, the R scores between the automated and human coding are now very high for the parties and issues and also quite high for the relations (0.79 and 0.70, respectively).

The final step of our evaluation is the Error Components Analysis. Tables 7 and 8 show the share of different errors with respect to the false negatives and false positives, respectively. More specifically, the errors were classified by an analysis of the original text source into different error categories. Such an analysis helps to improve the coding process by pointing to the most error-prone steps. All calculations are based on the gold standard test set for the 2003 electoral campaign.[30]

(Insert Table 7 about here)

With respect to the relation recognition, there are 114 false positives. 19.6 percent of them had an actor recognition problem, 42.0 percent showed an issue-recognition problem, and 38.4 percent had a problem with the recognition of both entities. Problems with the issue keywords in the gazetteer are the most important source of errors in the issue recognition.

---

standard nor the automated coding.

[30] We used the sentence-based method of relation detection for this analysis since the data are directly comparable to the gold standard using the sentence identifiers.

This means that the concept identification of the issues has some difficulties in resolving keyword ambiguities and false positives. The *out of context* problem is acute for both the parties and issues. *Out of context* means that there is neither a direct trigger for a keyword nor an anaphora in the sentence indicated by the gold standard. Therefore, we have no starting point for searching for a concept. This problem is a direct consequence of the difficulty that many semantic relations span more than one sentence. The high share of this error source adds evidence of the necessity to take more than one sentence into account when we establish relations. A relation recognition approach that takes this into account, e.g. the *±1 sentence* method, can, thus, handle this difficulty. The anaphora resolution with respect to the parties is an important source of error as well, since 44.6 percent of the errors in actor recognition stem either from a false or missing recognition of a referent. Here, more computational linguistic research is needed.

There are 28 false negatives in the relation detection of the 2003 test set (see Table 8). Half of the cases are caused by a gold-standard error, which means that the identified relations actually are correct, although they are not indicated by the gold standard. In other words, the inconsistency of the manual data also contributes to error in the measures used in all evaluations so far. Problems with our gazetteer and the anaphora cause some errors with respect to the false negatives, too. 14.3 percent of the errors, finally, stem from erroneously established relations.

From the Level of Measurement to the Level of Analysis

Assessing the quality of our relational data is only meaningful if these data are of actual use for analyses in the social sciences. Therefore, the value of the data at our level of analysis, i.e. election campaigns, has to be considered. In many studies on electoral competition, e.g., in expert surveys, party positions are measured once per electoral campaign. If more than one value is collected for party positions, they usually are aggregated

to the campaign level since most studies aim for comparisons among parties or different election campaigns. In the most general terms, data on relative frequencies thereby provide information on the weight of issues and parties in electoral campaigns and on the importance of specific issues for single parties. This means that various questions with respect to the study of electoral competition can be tested: theories of selective emphasis (Budge & Farlie, 1983), issue ownership (Petrocik, 2004), and attention shifts (Riker, 1986), for instance, all predominantly formulate expectations regarding the relative frequency of specific issues for parties. Some important approaches in political communication, such as agenda-setting and priming, in contrast, mainly elaborate hypotheses on how the media influences the visibility and public perception of actors and issues (see McCombs & Shaw, 1972; Behr & Iyengar, 1985; Kingdon, 1995). The exemplary analysis of the manually collected data on the Swiss 2003 and 2007 national electoral campaigns (our gold standard; see Figure 3a) and the automated data produced with the *simple distance* method (see Figure 3b) may allow a brief illustration of some benefits and shortcomings of data on relative frequencies in light of both research strands. The parties' relative weight in the election campaign is indicated by the size of the single network graphs. In addition, the parties are ordered according to their share of relations in the manual data set. The single-issue dots further illustrate the relative frequency with which an issue was used by the respective party.

(Insert Figure 3a about here)

(Insert Figure 3b about here)

With respect to the manual data, *Blick*'s coverage of the election campaign makes the Social Democrats the most important party, although it is only the second-biggest party with respect to its electoral strength. The biggest party, the populist right-wing Swiss People's Party (SVP), is ranked only third. This is, on the one hand, a confirmation of the impression that *Blick* is a newspaper with a slight bias to the left (Blum, 2005). On the other hand, the

result is counterfactual to the often-heard claims that boulevard media outlets strengthen populist parties with their scandalizing and personalizing style of reporting.

Figure 3 additionally shows the parties' issue emphases, which is indicated by the sizes of the issue dots. The Liberals, for instance, try to "own" economic issues by emphasizing budgetary rigor and economic liberalism. At the same time, they almost never speak of cultural issues such as cultural liberalism or anti-immigration viewpoints. This could be expected since the Liberals traditionally see their competence in economic policy-making. The SVP, on the contrary, focuses much more on cultural issues such as cultural liberalism. A slight emphasis on anti-immigration views can also be observed, an issue that all contestants seem to avoid. Such attention shifts toward usually taboo issues has been identified as an important element of the success of right-wing populist parties in Western Europe (see Kriesi et al., 2008). Other issues, such as welfare, are important for all parties. Here, positional data would be crucial in discerning the parties' characteristics and strategies, since it is unlikely that all parties share the same stance on the welfare regime. The value of data on relative frequencies, thus, has limits if there is no variation in the issues emphasized by the parties.

We have shown that our system is able to collect such data, too. Figure 3b, thus, presents the corresponding analysis of the automatically generated electoral data for both electoral campaigns.[31] Therefore, it allows for an overall discussion of the performance and problems of our approach. The order of the parties with respect to their relative frequency changes for the SVP and the Liberals, who now rank third. Compared to the manual data, the Liberals' relative frequency has decreased by 2.6 percent, while the share of the SVP has increased by 4.3 percent. The rank of the Social and Christian Democrats does not change,

[31] We take the frequency data generated by the simple distance' method since frequency data is directly comparable with commonly used manual CSA data.

and their differences in relative frequencies are also minor: 2.0 percent for the Social Democrats and 0.2 percent for the Christian Democrats. Thus, the rather high difference for the SVP was the main cause of the changes in the rank order. As we have discussed, this is due to our problems with anaphora resolution as well as the deficiencies of the manual data, as we have shown in the Error Components analysis.

With regard to the relative frequency of the issues, we first can notice several analogies between the graphs. First, welfare again is a very important issue for all parties. Further, the Liberals focus heavily on economic issues. Finally, we can also find the almost exclusive emphasis on the anti-immigration views by the SVP. However, there is also the odd result of the differing saliencies of economic liberalism, especially for the Social Democrats. Referring to our error components analysis, these problems most likely originate in deficiencies in our gazetteer. Thus, it seems that important keywords are missing or that there is an overlap between the keywords for economic liberalism and similar issues such as welfare. Overall, however, our automated relation recognition seems consistent in comparison with the manual data collection.

Future Prospects

Our system offers proof of feasibility such that computational linguistic tools should be implemented more commonly by political scientists to enhance their content analyses. Moreover, the results suggest that our automatic coding aligns parties and issues along the expected lines with only small differences. While not a big claim, this is, in fact, of greater interest to the political analyst because the results may be enough for other scholars to consider semi-automatic techniques for greater coverage and historical analysis. The CSA approach, meanwhile, has gained a solid reputation for delivering data on actor-issue relations (Kriesi et al. 2008, Bornschier, 2010, Helbling & Tresch, 2010, Helbling et al., 2010, Kleinnijenhuis, de Ridder, & Rietberg, 1997). Furthermore, research on implementations of the CSA ap-

proach has already been published for Dutch newspapers (van Atteveldt et al., 2008). That is, the CSA approach and our pipeline, basically, are applicable across other language corpora. What needs to be exchanged are the parser and tagger. Since the task to code CSA data is very similar to assessing predicate-argument structures, a common task in computational linguistics, these tools are available for many languages.

If the gazetteers are carefully adapted, our approach has the potential to collect data for diverse data collection tasks in political communication. For example, one could gather data on the campaign strategies of political parties or candidates, study long-term shifts in party positions, or assess the degree of political bias of different print media based on their reporting of electoral competition. Furthermore, the CSA approach is also not restricted to the issue positions of parties but can also be extended to trade unions, employer associations, and public administrative actors (Wueest, 2010).

The major effort for the annotators to perform analyses as described in this paper has to be made for the gazetteers. When we switched from human coding to automatic coding, we had to learn an important lesson: generic keywords are useful for human coders to grasp the relevant concepts, whereas, for automatic concept identification, they introduce a lot of noise. Therefore, it is necessary to extract the vocabulary actually used in the texts. We found that about one week was needed to establish, test, and update the gazetteers to be ready for the coding of the two election campaigns. We cannot be more precise with respect to effort-related questions beyond our experience of analyzing two election campaigns. Nevertheless, there are some basic guidelines that can be established for all campaign analyses. In general, if we proceeded to analyze more election campaigns, the effort would be dependent on the task. An adaptation of the gazetteers to analyze more newspapers for the same campaign would need only a little work. If the analysis were extended to other German-speaking countries or Swiss campaigns in other time periods, both the actor and issue gazetteers would need

a complete overhaul. At the least, the parties and politicians would change, and the keywords for the issues would certainly vary depending on the specific campaign context. For non-German speaking countries, parts of the implemented software also would need to be replaced. Additionally, the gazetteers would need to be built from scratch.

For the actors, i.e. the parties and politicians, comprehensive lists that can be transformed into gazetteers are readily available, e.g. from the Web sites of the parliamentary services of the countries under study. The establishment of the issue gazetteers is more time-consuming. However, here, tools to extract semantically similar words from the relevant corpus of newspaper articles are of great value to quickly establish lists containing keywords for specific issue categories. There are a lot of techniques that could be applied for this task, e.g. Latent Semantic Analysis (see Landauer, 1997, and http://marimba.d.umn.edu/cgi-bin/SC-cgi/index.cgi for a Web-based implementation) or co-occurrence analysis (see http://www.linguatools.de/disco/disco-gui.html for a Web-based implementation).

Currently recognizing issues is a binary decision that takes place regardless of how relevant or marginal an issue is in the whole document. Assessing and weighting the relevancy of an issue based on document-wide criteria might help to reduce the recognition of spurious issues.

Concerning our linguistic processing, several improvements could be addressed. In recent years, more broad-coverage parsers for German have become available: a shared-task parser evaluation on newspaper texts (Kübler, 2008) compares 3 different systems for phrase structure parsing, but the results do not indicate a clear-cut advantage either for the Stanford parser (Rafferty & Manning, 2008) or for the Berkeley parser (Petrov & Klein, 2008). For dependency parsing, the MaltParser (Hall & Nivre, 2008) was the only competitor; however, it shows nice properties of processing times even for long sentences. If we want to apply our

method of relation extraction to other newspapers, a better handling of long sentences is necessary.

Currently, our software is in an experimental state and requires close collaboration between social scientists and computational linguistics. For broader application, integration into a workbench for information extraction and text processing as GATE (Cunningham, 2002) is indispensable.

Further, negative or positive polarity orientation is the most important missing dimension in our automated relation detection. Such a polarity measure would allow us to gain more detailed insights from the data if it at least is able to capture whether a relationship is positive or negative. In the current stage of our project, we have to add polarity manually. For this task, a specifically designed Web application was builds.[32] The application allows simultaneous annotation by several coders and is equipped with an administrative panel to organize large-scale data collections, a database to store large amounts of text documents and annotated data, and a front end that allows the ergonomic annotation of CSA data. Concerning prospects to automatically recognize the polarity, different computational linguistic approaches from the realm of opinion-mining and sentiment detection are promising. First, some of the trigger keywords for the issues already express polarity though their connotation, e.g. the term "Scheininvalide" ("pseudo-handicapped" people) has a clear negative connotation with respect to welfare policies. Further, compositional sentiment detection based on polarity lexicons may be used to recognize whether statements are negative or positive (Klenner, Fahrni, & Petrakis, 2009). Sentences such as "*Bundesrat Deiss erteilt Peter Spuhlers Steuerabzug-Idee eine Abfuhr*" ("Bundesrat Deiss Rejects Peter

[32] For further information and a trial of the software framework, see http://www.bruno-wueest.ch/Software.html. All parts of the framework are open-source and, as long as third-party software is not involved, free to use for scientific purposes.

Spuhler's Proposition for Tax Reduction," *Blick*, August 28, 2003] illustrate that precise syntactic information and good set phrase recognition is needed, too. To gain deeper insight into the problem, we plan to collect the verbs of all sentences that have been annotated to date and compare their sub-categorization slots with the annotated entities of the core sentences (see Kim & Hovy 2004). However, as noted by (Pang & Lee, 2008), negative or positive sentiments can often be expressed subtly and are difficult to spot in isolated sentences or terms.

Discussion and Concluding Remarks

This paper presents an evaluation of a novel approach to (semi-)automatically collect relational data on electoral contests. The Core Sentence Analysis approach can, to a certain extent, be automated using computational linguistics tools and techniques. The automated production of data regarding the share of parties and their issue statements works quite well, although the data quality lags behind the output of manual annotations. If we consider only the occurrence of relations per article, our automated system produces decent results even in the context of reliability and external validity. The definition of a text passage of a few sentences as the unit of measurement momentarily offers the best balance between recall and precision. Entity identification works best if we consider a window of three sentences, and the relation recognition method that combines the nearest neighbors yields the best performance. The application of parsing methods substantively improves the precision of the relation recognition but still lags too far behind with respect to the recall.

Our evaluation, additionally, has shown further need to improve the software, linguistic rules, and gazetteers to make relation mining a widely accepted approach for social science content analyses, and, despite the prospect of speeding up the data-collection process in comparison with a purely manual approach, human coders are still heavily involved when it comes to the generation of a gazetteer and the recognition of polarity. There is, however, no

universal cure for the drawbacks of manual coding. In sum, our method, therefore, seems most useful for medium-n studies that collect data from several hundred or maybe thousands, but certainly not millions, of documents. In exchange, it produces fine-grained data and is able to do more than text classification. Furthermore, current methods are adapted to the Swiss context and the processing of the specific vocabulary used by boulevard newspapers. Consequentially, some future efforts will have to concentrate on the transfer of our methods to other newspapers, national settings, and languages.

References

Adams, J. F., Merrill, S., & Grofman, B. (2005). *A unified theory of party competition: A cross-national analysis integrating spatial and behavioral factors*. Cambridge, UK: Cambridge University Press.

Axelrod, R. (1976). *Structure of decision: The cognitive maps of political elites*. Princeton, NJ: Princeton University Press.

Behr, R. L., & Iyengar, S. (1985). Television news, real-world cues, and changes in the public agenda. *Public Opinion Quarterly, 49*, 38-57.

Blum, R. (2005). Politischer Journalismus in der Schweiz. In P. Donges (Ed.), *Politische Kommunikation in der Schweiz*. Bern, Switzerland: Haupt.

Bornshier, S. (2010). Cleavage *politics and the populist right*. The *new cultural conflict* in Western Europe. Philadelphia, PA: Temple University Press.

Brants, T. (2000). TnT – A statistical part-of-speech tagger. *Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000).* Seattle, WA, USA.

Budge, I., & Farlie, D. (1983). Party competition – Selective emphasis or direct confrontation? An alternative view with data. In H. Daalder & P. Mair (Eds.), *Western European party systems: Continuity and change* (pp. 267-305). London, UK: Sage.

Cortada, J. (2009). *How societies embrace information technology: Lessons for management and the rest of us*. Los Alamitos, CA: IEEE Computer Society.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, *36*, 223-254.

Downs, A. (1957). *An economic theory of democracy*. New York, NY: Harper Collins.

Evans, W. (2001). Computer environments for content analysis: Reconceptualizing the roles of humans and computers. In O. V. Burton (Ed.), *Computing in the Social Sciences and Humanities* (pp. 67-86). Urbana, IL: University of Illinois Press.

Foth, K., Daum, M., & Menzel, W. (2004). A broad coverage parser for German based on defeasible constraints. In H. Christiansen, P. R. Skadhauge & J. Villadsen (Eds.), *proceedings constraint solving and language processing, workshop proceedings, Datalogiske Skrifter No. 99* (pp. 88-101). Roskilde, Denmark: Roskilde Universitetscenter.

Foth, K., Menzel, W., & Schröder, I. (2005). Robust parsing with weighted constraints. *Natural Language Engineering, 11*(1), 1-25.

Foth, K. A. (2007). *Hybrid methods of natural language analysis*. Aachen, Country: Shaker.

Franzosi, R. (2004). *From words to numbers: Narrative, data, and social science*. Cambridge, UK: Cambridge University Press.

Fung, A., Graham, M., & Weil, D. (2007). Full *d*isclosure: The *perils and promise of transparency*. Cambridge, UK: Cambridge University Press.

Hall, J., & Nivre, J. (2008). A dependency-driven parser for German dependency and constituency representations. *Proceedings of the Workshop on Parsing German PaGe 08*, (June), 47-54. Association for Computational Linguistics. doi: 10.3115/1621401.1621408.

Helbling, M., Hoeglinger, D., & Wueest B. (2010). How political parties frame European integration. *European Journal of Political Research, 49*, 495-521.

Hillard, D., Purpura, S., & Wilkerson J. (2007). An active learning framework for classifying political text. *Paper presented at the 2007 Annual Meeting of the Midwest Political Science Association*, Chicago.

Hopkins, D., & King, G. (2010). Extracting systematic social science meaning from text. *American Journal of Political Science, 54*(1), 229-247.

Hug, S., & Schulz, T. (2007). Left-right Positions of Political Parties in Switzerland. *Party Politics* 13(3), 305-330.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of Coling 2004* (pp. 1367-1373). Geneva, Switzerland: COLING.

King, G., & Lowe, W. (2003). An automated tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization, 57*, 617-642.

Kingdon, J. W. (1995). *Agendas, alternatives and public policies*. Boston, MA: Little, Brown.

Kleinnijenhuis, J., de Ridder, J. A., & Rietberg, E. M. (1997). Reasoning in economic discourse: An application of the network approach to the Dutch press. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 191-209). Mahwah, NJ: Lawrence Erlbaum Associates.

Klenner, M., Fahrni, A., & Petrakis, S. (2009). PolArt: A robust tool for sentiment analysis. In K. Jokinen, & E. Bick (Eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009* (pp. 235-238). NEALT.

Klenner, M., & Ailloud, É. (2009). Optimization in coreference resolution is not needed: A nearly-optimal zero-one ILP algorithm with intentional constraints. *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*.

Koskeniemmi, K., & Haapalainen, M. (1996). GERTWOL – Lingsoft Oy. In R. Hausser (Ed.), *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994* (pp. 121-140). Tübingen, Germany: Niemeyer.

Kriesi, H., Grande E., Lachat, R., Dolezal M., Bornschier S., & Frey T. (2008). *West European politics in an age of globalization*. Cambridge, UK: Cambridge University Press.

Krippendorff, K. (2004). Content Analysis. An introduction to its methodology (2nd ed.). Thousand Oaks, CA: Sage.

Kübler, S. (2008). The PaGe 2008 shared task on parsing German. *Proceedings of the ACL2008 Workshop on Parsing German* (pp. 55-63). Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W08/W08-1008.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311-331.

Manning, C., & Schütze, H. (2002). *Foundations of statistical natural language processing*. Cambridge, UK: MIT Press.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of the mass media. *Public Opinion Quarterly,* 36, 176-187.

Meguid, B. M. (2005). Competition between unequals: The role of mainstream party strategy in niche party success. *American Political Science Review, 99*(3), 347-359.

Neuendorf, K. A., & Skalski, P. (2010). The content analysis guidebook online. Retrieved from the Cleveland State University Web site: http://academic.csuohio.edu/kneuendorf/content/

Osgood, C. E. (1956). Evaluative assertion analysis. *Litera, 3*, 47-102.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135. doi: 10.1561/1500000001.

Petrocik, J. R. (2004). Issue ownership and presidential campaigning, 1952-2000. *Political Science Quarterly, 118*(4), 599-626.

Petrov, S., & Klein, D. (2008). Parsing German with latent variable grammars. *Proc of ACL Parsing German Workshop* (pp. 33-39). Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W08/W08-1005.

Porter, B., Barker, K., & Hovy, E. H. (2007). Learning by reading: A prototype system, performance baseline and lessons learned. *Paper presented at the meeting of the Association for the Advancement of Artificial Intelligence.* Vancouver, Canada.

Rabinowitz, G., & Macdonald, E. S. (1989). A directional theory of voting. American Political Science Review*, 89*, 93-121.

Rafferty, A., & Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. *ACL08 Workshop on Parsing German* (pp. 40-46). Association for Computational Linguistics.

Riker, W. H. (1986). *The art of political manipulation*. New Haven, CT: Yale University Press.

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C., & Konstandi, O. (2007). Mining of functional relations between genes and proteins over biomedical scientific literature using a deep-linguistic approach. *Journal of Artificial Intelligence in Medicine*, *39*, 127-136. doi: 10.1016/j.artmed.2006.08.005.

Ruigrok, N., & van Atteveldt, W. (2007). Global angling with a local angle: How U.S., British, and Dutch newspapers frame global and local terrorist attacks. *The Harvard International Journal of Press/Politics, 12*, 68-90.

Scharl, A., & Weichselbraun, A. (2008). An automated approach to investigating the online media coverage of U.S. presidential elections. *Journal of Information Technology Politics*, *5*(1), 121-132.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Paper presented at the International Conference on New Methods in Language Processing.* Manchester, UK.

Schrodt, P. A. (2009). Reflections on the state of political methodology. *Newsletter of the Political Methodology Section, American Political Science Association, 17*(1), 1-4.

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political Science: KEDS – A program for the machine coding of event data. *Social Science Computer Review, 38*(4), 561-588.

Sennrich, R., Schneider, G., Volk M., & Warin M. (2009). A new hybrid dependency parser. *Proceedings of GSCL-Conference*. Potsdam.

van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2008). Parsing, semantic networks, and political authority. Using syntactic analysis to extract semantic relations from Dutch newspaper articles. *Political Analysis, 16*, 428-446.

West, M. D. (2001). *Applications of computer content analysis*. Westport, CT: Ablex Publishing.

Wittgenstein, L. (1984 [1921]). *Tractatus logico-philosophicus*. Frankfurt, Germany: Suhrkamp.

Wunsch, H. (2010). Rule-based and memory-based pronoun resolution for German: A comparison and assessment of data sources. *PhD thesis*, Universität Tübingen.

Wueest, B. (2010). Varieties of capitalist debates: How institutions shape public conflicts on economic liberalization in the U.K., France, and Germany. CIS Working Paper Nr. 64, 2010. *Center for Comparative and International Studies, ETH Zurich and University of Zurich*.

Zuell, C., & Landmann, J. (2005). Identifying national and international events using computer-assisted content analysis. *Paper presented at the first EASR conference in Barcelona*, Barcelona, Spain.

Author Note

We thank Hanspeter Kriesi, Michael Hess, Edgar Grande, Dominic Hoeglinger, Marc Helbling, Romain Lachat, Swen Hutter, Martin Dolezal, four anonymous reviewers, and the participants of the workshop titled “Computer- and Corpus-Linguistic Methods for Large-n Text Analysis in the Social Sciences” at the Freie Universität Berlin for their helpful comments. We wish to thank Lukas Rieder, Maël Mettler, and Laura Giess for their indispensable help with programming software and evaluating outputs. Finally, we thank the Swiss National Science Foundation as well as the Deutsche Forschungsgemeinschaft for their research support.

Bruno Wueest, MA.

University of Zurich

Fellow researcher at New York University and researcher at the Center for Comparative and International Studies, University of Zurich.

Dr. Simon Clematide

University of Zurich

Researcher at the Institute of Computational Linguistics, University of Zurich.

Alexandra Bünzli, lic. phil.

University of Zurich

Researcher at the Institute of Computational Linguistics, University of Zurich.

Daniel Laupper

University of Zurich

Student in political science at the Center for Comparative and International Studies, University of Zurich.

Timotheos Frey, lic. ès. sc. pol.
Secretary general of the Christian Democratic People's Party of Switzerland and former researcher at the Center for Comparative and International Studies, University of Zurich.

Correspondence concerning this article should be addressed to Bruno Wueest (wueest@ipz.uzh.ch).

Tables and Figures

Table 1: Example of a core sentence annotation

Die FDP ist ohne Wenn und Aber für Steuersenkungen. *[The FDP is without ifs and buts for tax reductions]* (*Blick*, October 4, 2003)		
Subject	**Polarity**	**Object**
FDP	*+ 1*	*tax reduction (meta-issue=economic liberalism)*

Figure 1: Software pipeline

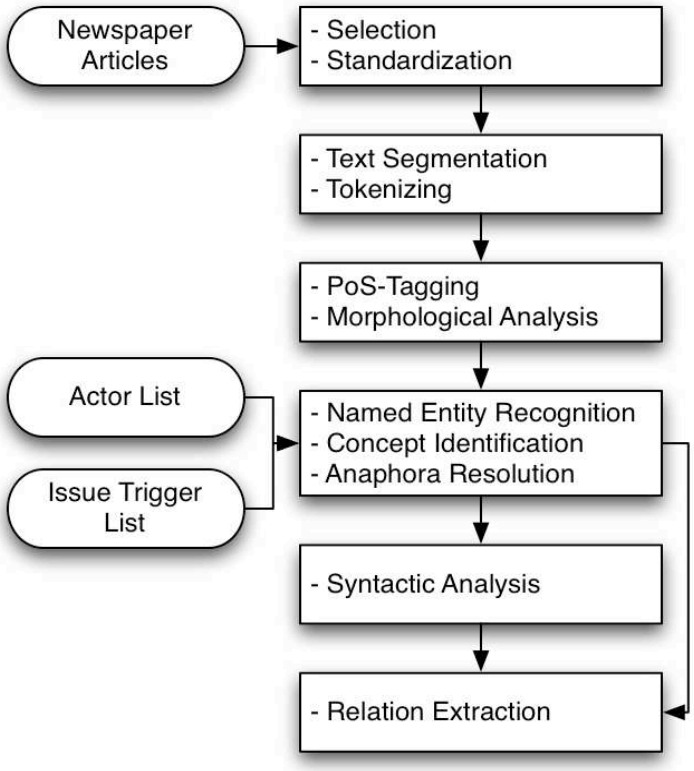


Figure 2: Parse trees

Figure 2a: Simple dependency tree

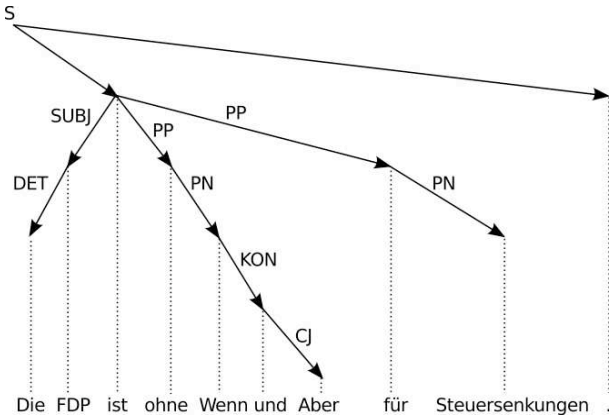
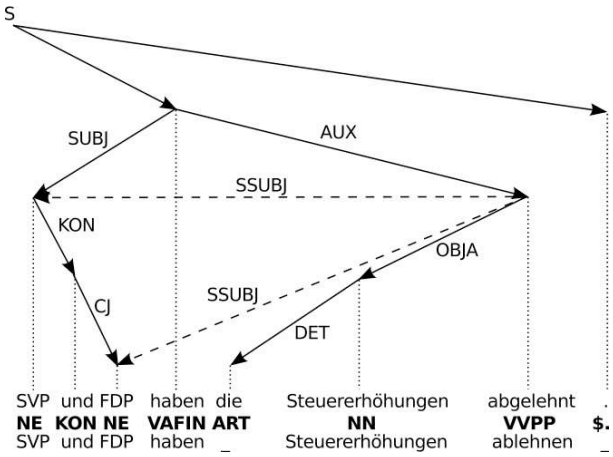


Figure 2b: Dependency tree with secondary relations



Notes: S=sentence, SUBJ=subject, DET=determiner, PP=prepositional phrase, PN=complement of prepositions, KON=conjunction, CJ=conjunct, AUX=auxiliary verb, OBJA=direct object, SSUBJ=semantic (or secondary) subject.

Table 2: Performance of the actor and issue recognition: frequencies per article

	Parties (N_G = 633)			***Issues (N_G = 633)***		
	article	*±1s*	*Sen-tence*	*article*	*±1s*	*sentence*
Development set						
Recall	0.85	0.64	0.39	0.80	0.61	0.38
Precision	0.23	0.49	0.69	0.35	0.55	0.69
F-score	0.36	0.55	0.49	0.48	0.58	0.49
N	2346	824	359	1458	699	350
Test set	***Parties (N_G = 238)***			***Issues (N_G = 238)***		
Recall	0.86	0.50	0.25	0.66	0.45	0.27
Precision	0.32	0.64	0.79	0.33	0.65	0.83
F-score	0.47	0.56	0.38	0.44	0.53	0.41
N	638	187	75	478	167	77

Table 3: Performance of the actor and issue recognition: occurrences per article

	Parties (N_G = 304)[1]			***Issues (N_G = 273)***		
	article	*±1s*[2]	*sentence*	*article*	*±1s*	*sentence*
Development set						
Recall	0.95	0.76	0.58	0.88	0.75	0.58
Precision	0.59	0.72	0.81	0.47	0.64	0.72
F-score	0.73	0.74	0.67	0.61	0.69	0.64
N	487	323	216	509	320	219
Test set	***Parties (N_G = 89)***			***Issues (N_G = 89)***		
Recall	0.98	0.7	0.47	0.8	0.64	0.46
Precision	0.57	0.76	0.82	0.43	0.68	0.82
F-score	0.72	0.73	0.6	0.56	0.66	0.59
N	153	82	51	165	84	50

Table 4: Performance of the relation detection: frequencies per article

	All combinations	**Simple distance**	**Subject filter**	**Parse tree distance**
Development set (N_G = 633)				
Recall	0.36	0.52	0.25	0.28
Precision	0.59	0.49	0.63	0.55
F-score	0.45	0.50	0.36	0.39
N	387	684	240	284
Test set (N_G = 238)				
Recall	0.23	0.38	0.13	0.15
Precision	0.68	0.53	0.69	0.63
F-score	0.34	0.44	0.22	0.24
N	81	170	45	57

Table 5: Performance of the relation detection: occurrences per article

	All combinations	**Simple distance**	**Subject filter**	**Parse tree distance**
Development set (NG = 633)				
Recall	0.47	0.59	0.34	0.37
Precision	0.63	0.59	0.68	0.67
F-score	0.54	0.59	0.45	0.48
N	291	398	194	219
Test set (NG = 238)				
Recall	0.36	0.48	0.21	0.24
Precision	0.70	0.60	0.70	0.64
F-score	0.47	0.53	0.33	0.35
N	61	97	37	45

Table 6: Inter-instrument reliability and external validity: comparison of the performance between manual and automated coding

	F-score	***Pearson's R***	
		level of articles	overall
Parties			
coder A vs. coder B	0.75	0.85	0.99
coder A vs. automated	0.56	0.57	0.97
coder B vs. automated	0.58	0.69	0.96
Issues			
coder A vs. coder B	0.67	0.75	0.83
coder A vs. automated	0.52	0.55	0.91
coder B vs. automated	0.45	0.59	0.71
Relations			
coder A vs. coder B	0.77	0.74	0.84
coder A vs. automated	0.63	0.50	0.79
coder B vs. automated	0.59	0.55	0.70

Notes: Only data from the 2003 test set are included in the calculations. With respect to the automated data, the method with the best fit from the previous evaluations was chosen.

Table 7: Error components analysis I: share of different sources in the false positives of sentence-based relation recognition in %

	Parties	**Issues**
Gazetteer	3.1	59.8
Out of context	47.7	33.5
Anaphora resolution	44.6	6.7
NER/rules	4.6	–
Total	100	100
N errors	65	90
Share of unique errors	19.6	42.0

Notes: Only the test set data for 2003 are considered in the calculations.

Table 8: Error components analysis II: share of different sources in the false negatives of sentence-based relation recognition

	Share in %
Gold standard error[1]	50
False negatives	
gazetteer (issues)	17.9
gazetteer (parties)	10.7
anaphora res (parties)	7.1
relation recognition	14.3
Total	100
N errors	28

Notes: Only the test set data for 2003 are considered in the calculations. [1] False negatives with respect to the manual coding.

Figure 3: Selective emphasis and relative frequency of parties in the Swiss national election campaigns 2003 and 2007:

Figure 3a: Manual data set

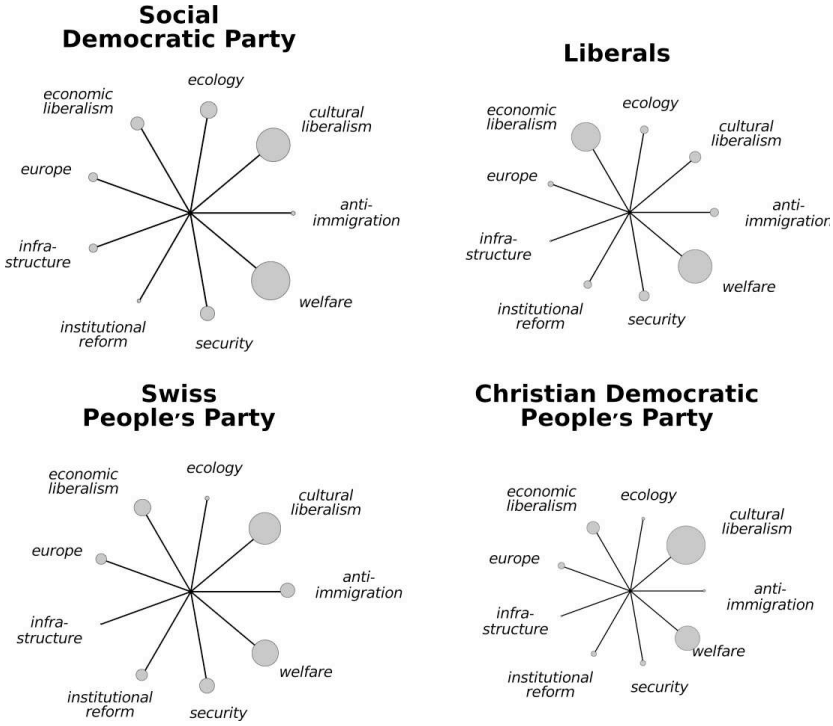
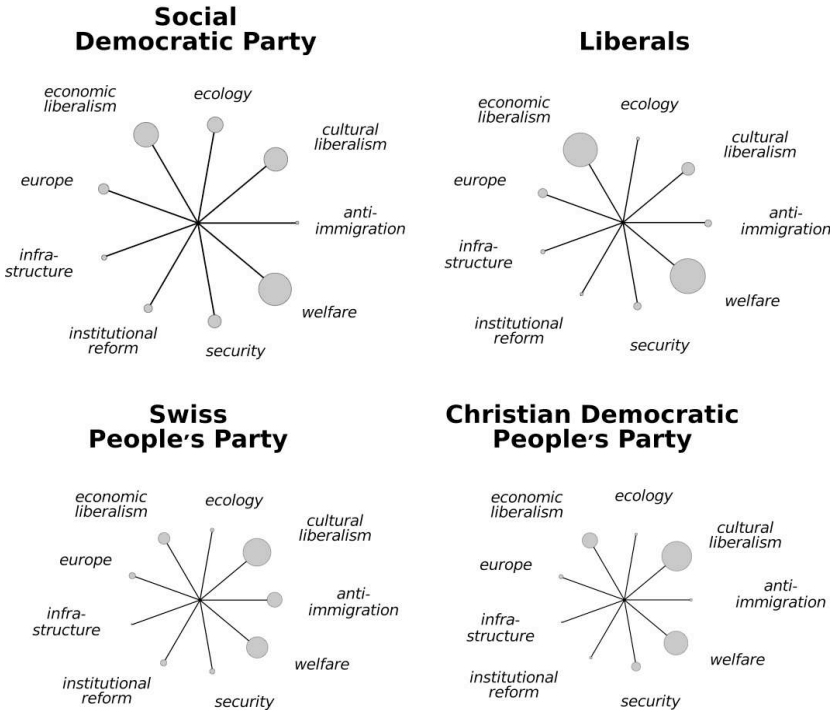


Figure 3b: Automated data set



Notes: Only parties with a share of more than 10 percent in all statements are included in the analysis. The sizes of the dots show the relative frequency of issues for the respective parties. The size of the single graphs indicates the relative frequency of the single parties in the electoral campaign. The saliencies for the manual data are SPS=29.6%, Liberals=25.1%; SVP=21.3%; and CVP=19.8%; the saliencies for the automated data are SPS=27.6%, Liberals=22.5%, SVP=25.6%, and CVP=19.6%.